УДК 681.324

П.Е. ПУСТОВОЙТОВ, канд. техн. наук,
ЭЛЬ САЕД АБДЕЛААЛ ЭЛЬ САЕД МОХАМЕД

МЕТОДИКА АНАЛИЗА МНОГОКОМПОНЕНТНЫХ ВХОДНЫХ ПОТОКОВ В КОМПЬЮТЕРНЫХ СЕТЯХ

Запропоновано методику розщеплення композиційного потоку даних, що поступає на вхід вузла комп'ютерної мережі на складові. Процедура розщеплення основана на мінімізації сумарної помилки ідентифікації елементарних потоків.

It was suggested the methodic of composition data flow separating, which is received by network switching node, to rectangular components. The procedure of separation bases on summary error minimization in elementary flows identification.

Анализ литературы. В реальных компьютерных сетях (КС) поток заявок, поступающих на вход узла сети, является суперпозицией нескольких потоков, отличающихся друг от друга численными значениями своих характеристик (интенсивностью, объемом пакета, законом распределения интервала между заявками и т.д.). Так, например, гистограмма, описывающая случайную величину объема пакета, полученная по данным потока заявок на одном из узлов компьютерной сети НТУ «ХПИ» в случайно выбранный рабочий день 12.09.2006 в интервале $11^{00}-12^{00}$ имеет вид, приведенный на рис. 1. Для каждой заявки этого потока известны моменты поступления и окончания обслуживания.

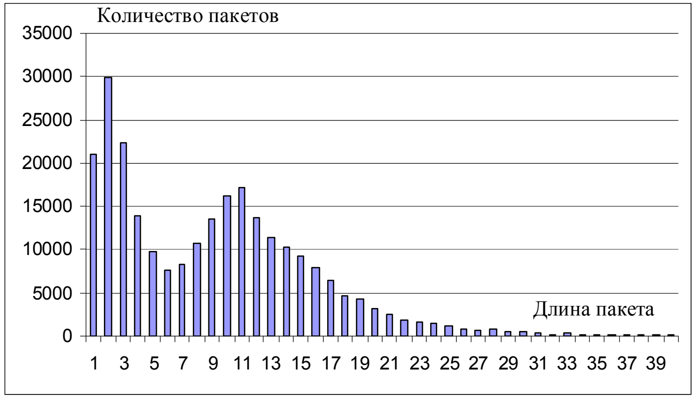

Рис. 1. Гистограмма случайной длины пакета

Из рисунка видно, что поток заявок представляет собой суперпозицию двух потоков с разными плотностями распределения длины пакета. Отметим, что в известной литературе, например, в [1 – 3] и других по анализу и оценке эффективности сетей, на вход которых поступает многокомпонентный поток, учитываются только различия в длине интервалов между пакетами в предположении, что входящий поток – марковский.

При этом различия в длине пакетов не рассматривались. В связи с этим, **целью статьи** является разработка методики исследования многокомпонентного потока на входе КС с учетом различий в основных характеристиках составляющих потока.

Непосредственный анализ эффективности функционирования узла КС для такого типа входных потоков затруднителен ввиду отсутствия их удовлетворительных аналитических описаний. Поставим задачу расщепления наблюдаемого результирующего потока на элементарные составляющие. Рассмотрим технологию решения этой задачи для реального потока с гистограммой случайной длины пакета, представленной на рис. 1.

Предположим, что этот поток составлен из двух потоков с релеевским распределением длины пакета, параметры которых различны. Заявки из этих потоков поступают на вход узла с соответствующими вероятностями.

Постановка задач исследования.

1. Найти параметры распределений длины пакета для составляющих результирующего потока.
2. Найти вероятности появления заявок для этих потоков.
3. Проверить справедливость гипотез относительно законов распределений длин пакетов.
4. Обосновать методику собственно расщепления результирующего потока на составляющие.
5. Найти плотности распределения продолжительности интервалов между заявками для составляющих результирующего потока и оценить статистически их параметры.

Основные результаты.

Введем

$f_1(x)=\dfrac{x}{\sigma_1^{\,2}}e^{-\frac{x^2}{2\sigma_1^2}}$ – плотность распределения длины пакета для первой составляющей суммарного потока;

$f_2(x)=\dfrac{x}{\sigma_2^{\,2}}e^{-\frac{x^2}{2\sigma_2^2}}$ $(\sigma_2>\sigma_1)$ – плотность распределения длины пакета для второй составляющей суммарного потока;

p – вероятность появления сообщения, принадлежащего первому потоку;

$q=1-p$ – вероятность появления сообщения, принадлежащего второму потоку;

y_j – относительная частота появления пакета, длина которого x_j соответствует j-му подынтервалу интервала наблюдаемых длин пакетов, $j = 1, 2, ..., n$.

Неизвестные параметры σ_1, σ_2, p найдем, используя метод наименьших квадратов [4, 5]

$$J = \sum_{j=1}^{n}\left[p\frac{x_j}{\sigma_1^2}e^{-\frac{x_j^2}{2\sigma_1^2}} + (1-p)\frac{x_j}{\sigma_2^2}e^{-\frac{x_j^2}{2\sigma_2^2}} - y_j\right]^2 . \qquad (1)$$

Составим уравнения для отыскания σ_1, σ_2, p. Имеем

$$\frac{\partial J}{\partial p} = 2\sum_{j=1}^{n}\left[p\frac{x_j}{\sigma_1^2}e^{-\frac{x_j^2}{2\sigma_1^2}} + (1-p)\frac{x_j}{\sigma_2^2}e^{-\frac{x_j^2}{2\sigma_2^2}} - y_j\right]\left(\frac{x_j}{\sigma_1^2}e^{-\frac{x_j^2}{2\sigma_1^2}} - \frac{x_j}{\sigma_2^2}e^{-\frac{x_j^2}{2\sigma_2^2}}\right) = 0 .$$

Отсюда

$$p = \frac{\sum_{j=1}^{n}\left(y_j - \frac{x_j}{\sigma_2^2}e^{-\frac{x_j^2}{2\sigma_2^2}}\right)\left(\frac{x_j}{\sigma_1^2}e^{-\frac{x_j^2}{2\sigma_1^2}} - \frac{x_j}{\sigma_2^2}e^{-\frac{x_j^2}{2\sigma_2^2}}\right)}{\sum_{j=1}^{n}\left(\frac{x_j}{\sigma_1^2}e^{-\frac{x_j^2}{2\sigma_1^2}} - \frac{x_j}{\sigma_2^2}e^{-\frac{x_j^2}{2\sigma_2^2}}\right)} . \qquad (2)$$

Далее

$$\frac{\partial J}{\partial \sigma_1} = \frac{2p}{\sigma_1^3}\sum_{j=1}^{n}\left[p\frac{x_j}{\sigma_1^2}e^{-\frac{x_j^2}{2\sigma_1^2}} + (1-p)\frac{x_j}{\sigma_2^2}e^{-\frac{x_j^2}{2\sigma_2^2}} - y_j\right]\left(\frac{x_j}{\sigma_1^2} - 2\right)x_j e^{-\frac{x_j^2}{2\sigma_1^2}} = 0 ;$$

$$\frac{\partial J}{\partial \sigma_2} = \frac{2(1-p)}{\sigma_2^3}\sum_{j=1}^{n}\left[p\frac{x_j}{\sigma_1^2}e^{-\frac{x_j^2}{2\sigma_1^2}} + (1-p)\frac{x_j}{\sigma_2^2}e^{-\frac{x_j^2}{2\sigma_2^2}} - y_j\right]\left(\frac{x_j}{\sigma_2^2} - 2\right)x_j e^{-\frac{x_j^2}{2\sigma_2^2}} = 0 .$$

Отсюда

$$\sum_{j=1}^{n}\left[p\frac{x_j}{\sigma_1^2}e^{-\frac{x_j^2}{2\sigma_1^2}} + (1-p)\frac{x_j}{\sigma_2^2}e^{-\frac{x_j^2}{2\sigma_2^2}} - y_j\right]\left(\frac{x_j}{\sigma_1^2} - 2\right)x_j e^{-\frac{x_j^2}{2\sigma_1^2}} = 0 ; \qquad (3)$$

$$\sum_{j=1}^{n}\left[p\frac{x_j}{\sigma_1^2}e^{-\frac{x_j^2}{2\sigma_1^2}} + (1-p)\frac{x_j}{\sigma_2^2}e^{-\frac{x_j^2}{2\sigma_2^2}} - y_j\right]\left(\frac{x_j}{\sigma_2^2} - 2\right)x_j e^{-\frac{x_j^2}{2\sigma_2^2}} = 0 . \qquad (4)$$

Для численного решения системы уравнений (2) – (4) удобно использовать следующую итерационную процедуру.

Сначала рассчитаем начальные приближения к искомому решению. С этой целью используем то обстоятельство, что

$$x^* = \arg\max_x \frac{x}{\sigma^2}e^{-\frac{x^2}{2\sigma^2}} = \sigma ; \qquad (5)$$

$$f(x^*) = \frac{1}{\sigma}e^{-0.5} . \qquad (6)$$

Тогда из рис. 1 имеем $\sigma_1^{(0)} = 1$, $\sigma_2^{(0)} = 5{,}5$, $p^{(0)} = 0{,}32$.

Далее, задав $\sigma_1^{(0)}$, $\sigma_2^{(0)}$ в (2), рассчитаем $p^{(1)}$. Теперь, подставив $p^{(1)}$ в (3) и (4), получим систему уравнений относительно σ_1 и σ_2, решив которую, найдем $\sigma_1^{(1)}$, $\sigma_2^{(1)}$. Эти значения используем для уточнения p и т.д. В результате реализации этой процедуры, получим $\sigma_1^* = 0{,}892$, $\sigma_2^* = 5{,}34$, $p^* = 0{,}38$.

Для проверки справедливости принятой гипотезы о плотностях распределения длин пакетов составляющих результирующего потока используем критерий χ^2 [6, 7]. В соответствии с этим критерием вычислим

$$\chi^2_{\text{набл}} = \sum_{j=1}^{n}\frac{[f_\Sigma(x_j)\Delta - y_j]^2 n}{f_\Sigma(x_j)\Delta} , \qquad (7)$$

где $f_\Sigma(x_j) = p\frac{x_j}{\sigma_1^2}e^{-\frac{x_1^2}{2\sigma_1^2}} + (1-p)\frac{x_j}{\sigma_2^2}e^{-\frac{x_1^2}{2\sigma_2^2}}$;

$\Delta = \frac{x_{\max}}{n}$.

Непосредственные вычисления дают $\chi^2_{\text{набл}} = 31{,}4$. Критическое значение критерия для числа степеней свободы, превышающего 30, и уровня значимости $\alpha = 0{,}005$ $\chi^2_{\text{набл}} = 43{,}77$. Так как $\chi^2_{\text{набл}} < \chi^2_{\text{кр}}$, делаем вывод, что принятые гипотезы не противоречат опытным данным.

Перейдём, наконец, к задаче собственно расщепления смеси. Смысл задачи состоит в том, чтобы выработать некоторое решающее правило [8], позволяющее отнести каждый конкретный пакет к одному из потоков, составляющих результирующий поток. Наблюдаемый параметр пакета – его длина. Введём гипотезу H_0 – пакет принадлежит первому потоку и альтернативную гипотезу H_1 – пакет принадлежит второму потоку. Введём рандомизированное решающее правило $A(x)$, состоящее в том, что если наблюдаемый параметр имеет значение x, то решение о справедливости гипотезы H_0 отвергается с вероятностью $A(x)$. Тогда уровень значимости критерия [9] (вероятность отвергнуть H_0, когда она верна) будет равен

$$\int_{\Omega} f\left(x/H_0\right)A(x)dx = \alpha , \tag{8}$$

а мощность критерия (вероятность отвергнуть H_0, когда верна гипотеза H_1) запишется в виде

$$\int_{\Omega} f\left(x/H_1\right)A(x)dx = \mu . \tag{9}$$

Здесь Ω – множество возможных значений x.

Выберем решающее правило таким образом, чтобы максимизировать мощность критерия (9) при заданном уровне значимости (8). Полученная задача является задачей континуального линейного программирования [10]. Её решение имеет вид:

$$A^*(x) = \begin{cases} 1, \text{ если } \left(f\left(x/H_1\right) / f\left(x/H_0\right)\right) > \lambda, \\ 0 \text{ в противном случае,} \end{cases} \tag{10}$$

где λ является решением уравнения

$$\int_{\tau_\lambda} f\left(x/H_0\right)dx = \alpha , \qquad \tau_\lambda = \left\{ x : f\left(x/H_1\right) > \lambda f\left(x/H_0\right)\right\}. \tag{11}$$

Решим неравенство

$$\frac{f\left(x/H_1\right)}{f\left(x/H_0\right)} = \frac{\sigma_1^2}{\sigma_2^2}\exp\left\{-\frac{x^2}{2}\left(\frac{1}{\sigma_2^2}-\frac{1}{\sigma_1^2}\right)\right. > \lambda .$$

Отсюда

$$\frac{x^2}{2}\left(\frac{\sigma_2^2-\sigma_1^2}{\sigma_1^2\sigma_2^2}\right) > \ln\frac{\lambda\sigma_2^2}{\sigma_1^2},$$

$$x^2 > \frac{2\sigma_1^2\sigma_2^2}{\sigma_2^2-\sigma_1^2}\ln\frac{\lambda\sigma_2^2}{\sigma_1^2} .$$

Так как искомое значение $x > 0$, то

$$x > \left(\frac{2\sigma_1^2\sigma_2^2}{\sigma_2^2-\sigma_1^2}\ln\frac{\lambda\sigma_2^2}{\sigma_1^2}\right) = \alpha . \tag{12}$$

Теперь, используя (12), найдем пороговое значение x, задающее решающее правило,

$$\int_{\alpha}^{\infty}\frac{x}{\sigma_1^2}e^{-\frac{x^2}{2\sigma_1^2}}dx = e^{-\frac{x^2}{2\sigma_1^2}} = \alpha , \tag{13}$$

откуда, с учетом (12),

$$\frac{2\sigma_1^2\sigma_2^2}{\sigma_2^2-\sigma_1^2}\ln\frac{\lambda\sigma_2^2}{\sigma_1^2} = -2\sigma_1^2\ln\alpha . \tag{14}$$

Подставляя (14) в (12), получим

$$x^* = \left(-2\sigma_1^2\ln\alpha\right)^{\frac{1}{2}} . \tag{15}$$

Таким образом, задавая вероятность ошибки первого рода, можно рассчитать пороговое значение длины пакета, при превышении которого пакет должен быть отнесён ко второму потоку, а в противном случае – к первому. К сожалению, при использовании описанного критерия вероятность ошибки второго рода (отвергнуть гипотезу H_1, когда она верна) однозначно навязывается выбранной вероятностью ошибки первого рода. В связи с этим в теории статистических решений часто используется другое решающее правило, основанное на так называемом критерии «идеального наблюдателя» [6], при котором минимизируется сумма вероятностей ошибок первого и второго рода. Легко показать, что пороговое значение при этом определяется абсциссой точки пересечения $f\left(x/H_0\right)$ и $f\left(x/H_1\right)$. Найдём это пороговое значение для модели (9). Имеем

$$\frac{x}{\sigma_1^2}e^{-\frac{x^2}{2\sigma_1^2}} = \frac{x}{\sigma_2^2}e^{-\frac{x^2}{2\sigma_2^2}} ,$$

откуда

$$\ln\frac{\sigma_2^2}{\sigma_1^2} = \frac{x^2}{2}\left(\frac{1}{\sigma_1^2}-\frac{1}{\sigma_2^2}\right);$$

$$x^* = \sigma_1\sigma_2 \left(\frac{2\ln\frac{\sigma_2^{\,2}}{\sigma_1^{\,2}}}{\sigma_2^{\,2}-\sigma_1^{\,2}} \right)^{\frac{1}{2}} . \qquad (16)$$

При использовании порога (16) значения уровня значимости и мощности критерия будут соответственно равны

$$\alpha = \int\limits_{x'}^{\infty} f\left(x \big/ H_0 \right) dx = e^{-\frac{(x')^2}{2\sigma_1^{\,2}}} ; \qquad (17)$$

$$\mu = \int\limits_{x'}^{\infty} f\left(x \big/ H_1 \right) dx = e^{-\frac{(x')^2}{2\sigma_2^{\,2}}} . \qquad (18)$$

Задача расщепления результирующего потока на составляющие решена. Далее для каждой из выделенных составляющих потока были построены гистограммы случайной длины интервалов между сообщениями. Ниже приведены эти гистограммы для первого и второго потоков, полученных при расщеплении с использованием порогового значения (16).

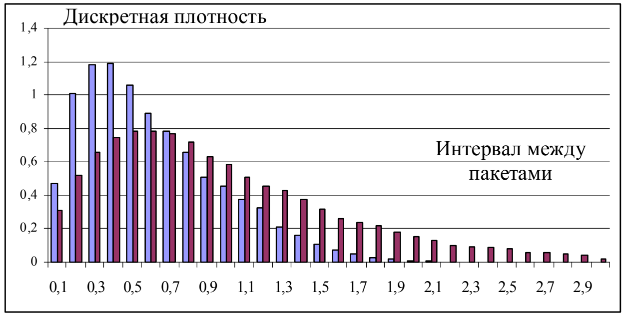

Рис. 2. Гистограммы распределения интервала между пакетами для первого и второго потоков

Аппроксимируем полученные гистограммы распределениями Эрланга второго рода с плотностью распределения

$$\varphi(t) = \lambda(\lambda t)e^{-\lambda t} .$$

Неизвестный параметр λ найдём методом максимума правдоподобия [13]. Введём множества E_1 и E_2, составленные из номеров пакетов, отнесённых после расщепления соответственно к первому и второму потокам.

Сформируем функции правдоподобия для отыскания соответствующих интенсивностей потоков λ_1 и λ_2:

$$P_1 = \prod_{j\in E_1} \lambda_1(\lambda_1 t_j)e^{-\lambda_1 t_j} , \; P_2 = \prod_{j\in E_2} \lambda_2(\lambda_2 t_j)e^{-\lambda_2 t_j} .$$

Логарифмические функции правдоподобия имеют вид:

$$L_1 = 2m_1 \ln\lambda_1 + \sum_{j\in E_1} (\ln t_j - \lambda_1 t_j) ;$$

$$L_2 = 2m_2 \ln\lambda_2 + \sum_{j\in E_2} (\ln t_j - \lambda_2 t_j) ,$$

где m_1 и m_2, соответственно, есть число пакетов в E_1 и E_2.

Далее имеем $\dfrac{dL_1}{d\lambda_1} = \dfrac{2m_1}{\lambda_1} - \sum\limits_{j\in E_1} t_j = 0$; $\dfrac{dL_2}{d\lambda_2} = \dfrac{2m_2}{\lambda_2} - \sum\limits_{j\in E_2} t_j = 0$.

Отсюда

$$\lambda_1 = \frac{2m_1}{\sum\limits_{j\in E_1} t_j}, \quad \lambda_2 = \frac{2m_2}{\sum\limits_{j\in E_2} t_j} . \qquad (19)$$

Полученные значения интенсивностей потоков могут быть теперь использованы для построения полумарковских моделей функционирования узла КС.

Выводы. Таким образом, предложена и обоснована методика расщепления потока данных, поступающих на вход узла КС, на составляющие результирующий поток элементарные потоки. Процедура расщепления основана на применении статистического решающего правила, минимизирующего суммарную ошибку идентификации элементарных потоков. Полученные результаты могут быть использованы для построения полумарковских моделей КС.

Список литературы: 1. *Столлинг В.* Современные компьютерные сети. – СПб.: Питер, 2003. – 783 с. **2.** *Крылов В.В., Самохвалова С.С.* Теория телетрафика и ее приложения. – СПб.: БХВ Петербург, 2005. – 288 с. **3.** *Таненбаум Э.* Компьютерные сети. – СПб.: Питер, 2003. – 992 с. **4.** *Линник Ю.В.* Метод наименьших квадратов и основы теории обработки наблюдений. – М.: Физматгиз, 1962. – 352 с. **5.** *Мудров В.И., Кушко В.Л.* Методы обработки измерений. – М.: Радио и связь, 1983. – 304 с. **6.** *Крамер Г.* Математические методы статистики. – М.: Мир, 1975. – 648 с. **7.** *Вентцель Е.С.* Теория вероятностей. – М.: Наука, 1964. – 564 с. **8.** *Леман Э.* Проверка статистических гипотез. – М.: Наука, 1964. – 534 с. **9.** *Вальд А.* Статистические решающие функции. – М.: Наука, 1967. – 384 с. **10.** *Раскин Л.Г., Кириченко И.О.* Континуальное линейное программирование. – Х.: ВИВВ, 2005. – 176 с. **11.** *Ван Трис.* Теория обнаружения, оценок и модуляции. – М.: Сов. радио, 1972. – 744 с. **12.** *Ниммельблау Д.* Анализ процессов статистическими методами. – М.: Мир, 1973. – 957 с. **13.** *Ван дер Варден Б.Л.* Математическая статистика. – М.: ИЛ, 1960. – 434 с.

Поступила в редакцию 13.10.2006